HR EMPLOYEE ATTRITION

**1. Problem Definition**

Employee Attrition (also known as "employee churn") is a very costly problem for companies. This means how many employees would leave the company and how it might affect the companies. The actual cost of replacing an employee can be quite high.

A study by the Center for American Progress found that companies typically pay about one fifth of an employee's salary to replace that one, and the cost could rise dramatically if the highest paid employees or employees had to be replaced.

In other words, the cost of replacing multiple employers remains significant. This is because of the time spent negotiating and finding a replacement, signing bonuses, and product losses for a few months while a new employee becomes accustomed to a new role.

Understanding why and when employees are likely to leave can lead to actions to improve staff retention and to plan for new recruitment in advance. I will use a systematic step-by-step method using a method that can be used for a variety of ML problems. This project will fall under what is best known as HR Analytics or People Analytics.

Here I have tried to find out the answers to:

How likely is it for an employee to leave the company?

What external factors might influence an employee to churn out?

What might be the indicators for an employee to leave the company?

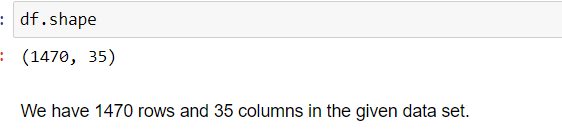
What steps could be taken by the companies to retain their employees?

**2.Data Analysis**

The data I have analysed is from the IBM HR Analytics dataset sourced from their website.

**2.1 About Data:**

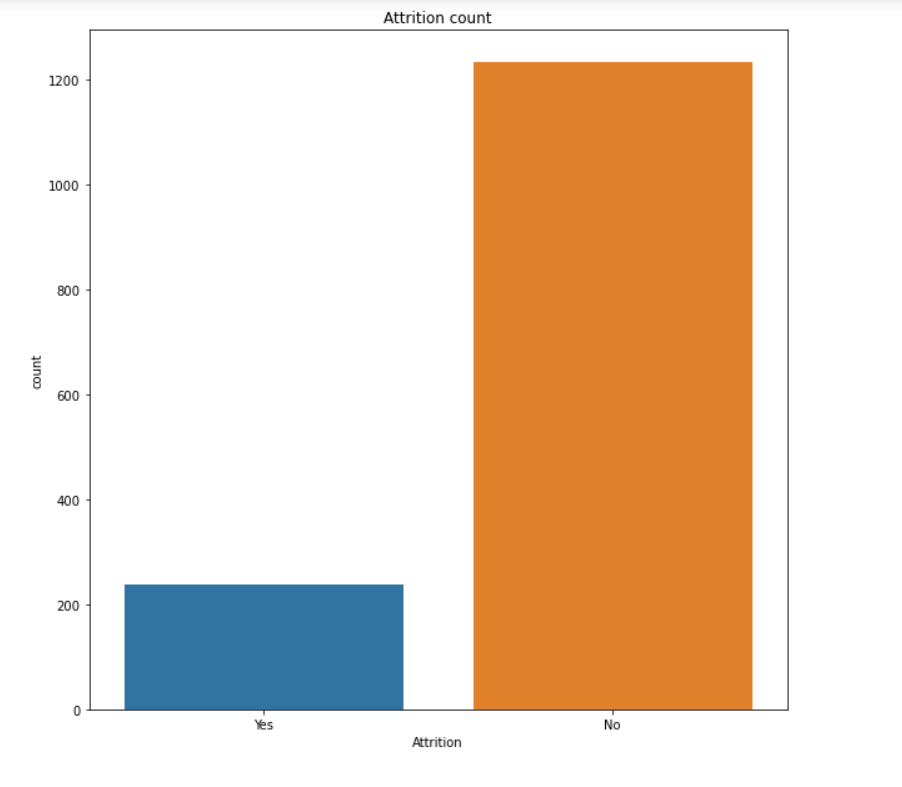
The dataset consists of 1470 employee details with 35 columns consisting various features.



The below columns from the dataset acts as features on which we can study how much it affects our target column which is attrition.

[Age,BusinessTravel, DailyRate ,Department, DistanceFromHome ,Education EducationField ,EmployeeCount ,EmployeeNumber ,EnvironmentSatisfaction ,Gender ,HourlyRate ,JobInvolvement ,JobLevel ,JobRole ,JobSatisfaction ,MaritalStatus ,MonthlyIncome ,MonthlyRate ,NumCompaniesWorked ,Over18 ,OverTime ,PercentSalaryHike ,PerformanceRating ,RelationshipSatisfaction ,StandardHours ,StockOptionLevel ,TotalWorkingYears ,TrainingTimesLastYear ,WorkLifeBalance ,YearsAtCompany ,YearsInCurrentRole ,YearsSinceLastPromotion ,YearsWithCurrManager]

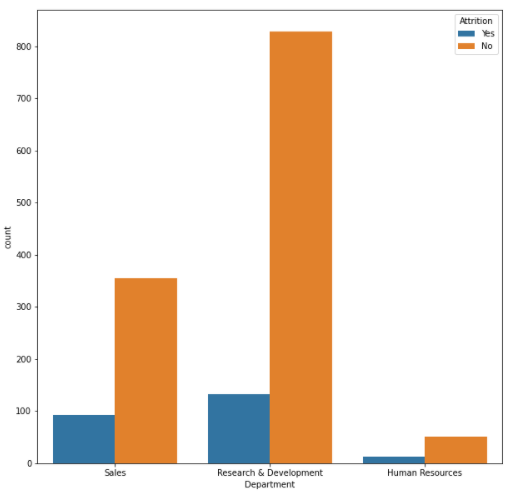
The dataset has no missing values and we have two unique values for the attrition column which is our target column which is yes or no.

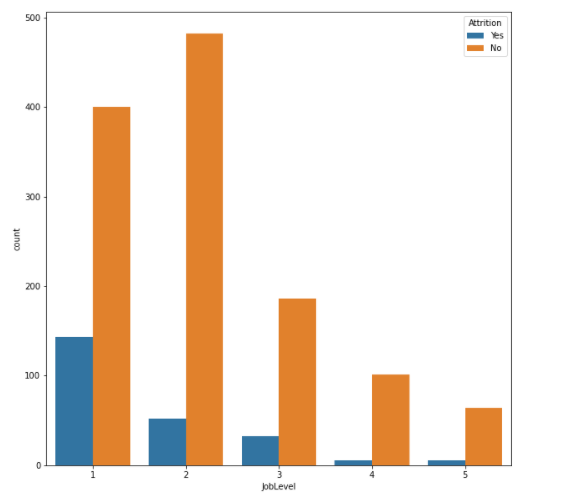


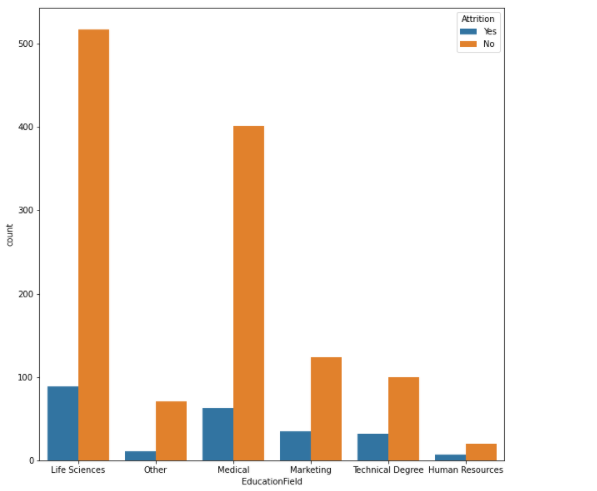
As we can see from above that the count plot for attrition column, that in the dataset there are a few employees who have left the company, and this would make our target class imbalanced which might result in the ML model giving more weightage for the output being No.

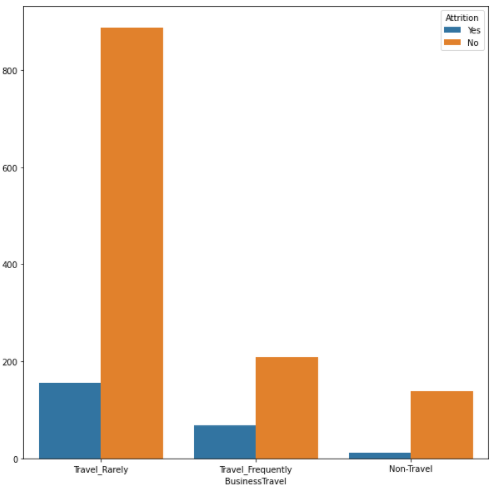
Below are the trends we can see for all features compared with attrition so we can have a better idea on how the features are impacting the attrition of employees.

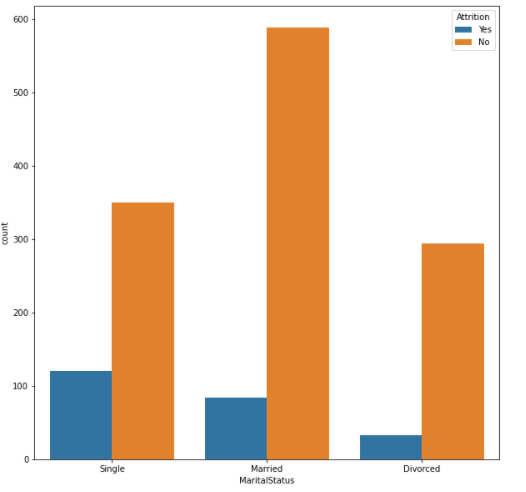
Machine learning algorithms usually work best when the number of situations for each class is almost equal. We will have to deal with this aspect of inequality before using our machine learning algorithms.

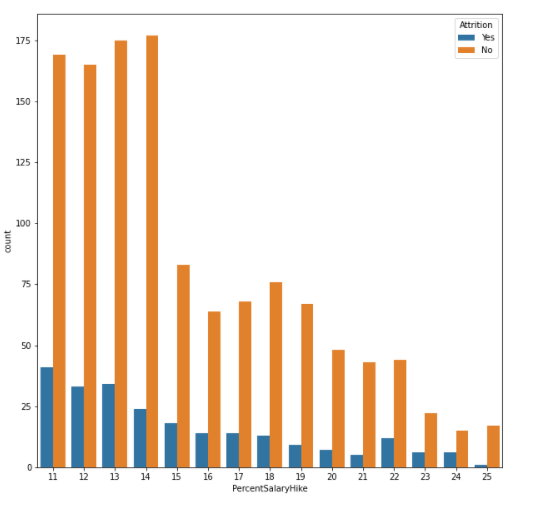


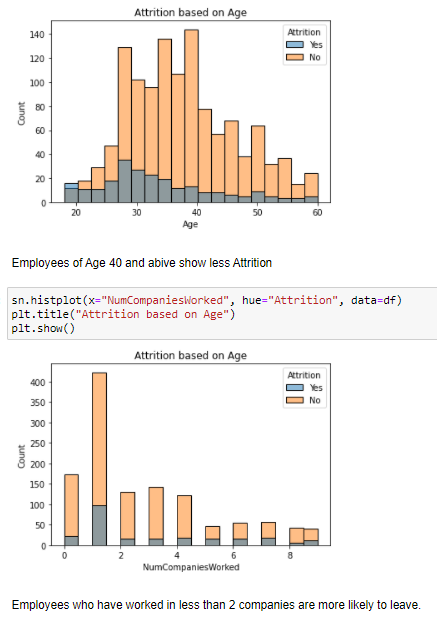












We can see based on Department:

Employees under sales are more likely to leave

Employees under Research and Development are less likely to leave.

We can see based on Job levels:

There are more people in Job level 1 or 2

Employees under 1,2 and 3 are more likely to leave

Employees above level 3 are less likely to leave.

We can see based on Education Field:

There are more people from Life Sciences and Medical.

Employees from HR, Marketing and Technical degree are more likely to leave

Employees from other fields are relatively less likely to leave.

We can see based on Business Travels:

There are more people who travel rarely.

Employees who travel frequently are more likely to leave

Employees who are non-travellers are less likely to leave.

We can see based on Business Travels:

There are more people who are married.

Employees who are single are more likely to leave

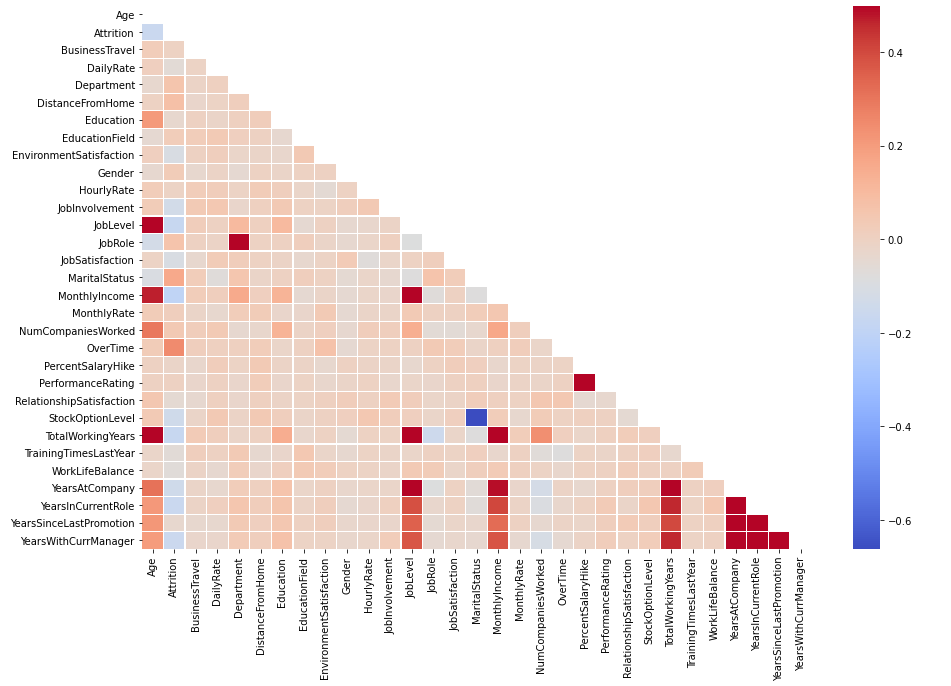
Employees who are married and divorced are less likely to leave.

We can see based on Salary hike that employees with salary hike more than 17% are more likely to stay.

We can see based on Age that employees above 35 are more likely to stay.

**Correlation**

Let’s take a look at some of the most notable connections. It is worth remembering that integration coefficients only measure formal integration.



We see that the Attrition is positively related to Over time and Marital Status.

and is negatively correlated to Job level, Monthly income and Total working years.

# **3. EDA Concluding Remarks**

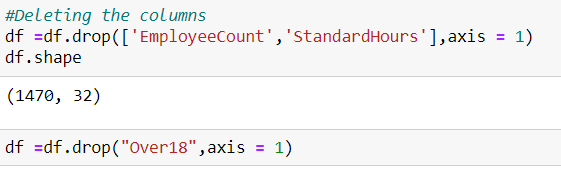
* Database does not include any missing or incorrect data values, and all features are the correct data type.
* The strongest relationships with targeted factors are: Performance Rate, Monthly Rate, Num Companies Working, Distance from Home.
* The worst relationships with targeted factors are: Total Working Years, Job Level, Current Rates, and Monthly Benefits.
* Individual employees account for the highest number of retirees, compared to their married and divorced counterparts.
* Few people leave when they reach their 2 years of company.
* People who live far away from their work show a higher rate of abandonment compared to their counterparts.
* People who regularly travel show a higher rate of leaving compared to others.
* People who have to work overtime show a higher rate of resignation compared to their counterparts.
* Employees who have previously worked for several companies show a higher rate of leave than their counterparts.
* Employees with salary hike more than 17% are more likely to stay.

**4. Pre-Processing Pipeline**

**4.1 Dropping columns**



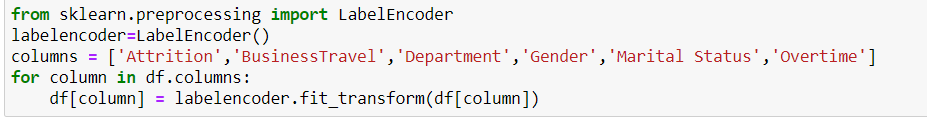
I dropped the “EmployeeNumber” column as it had unique number for each employee thereby having no impact on our target variable “Attrition” as it does not have anything to do with employee churn.

Dropping

* EmployeeCount as the value is ‘1’ for all the rows.
* StandardHours as the value is 80 for all the rows.
* Over18 as the value is Y(“yes”) for all the rows all employees in the dataset are above 18 years.

As all these columns have the same value for all employees, it will not be correlated to the employee attrition and hence I have removed the four columns.

**4.2 Encoding**

I have encoded the categorical columns which have 2 or 3 unique values using label encoder.

Encoding is necessary for the categorical variables as the ML algorithm cannot interpret object type data.



I have encoded the remaining the remaining columns with one hot encoder.

**4.3 Balancing the target class**

As we had seen earlier that the class Attrition was not balanced.

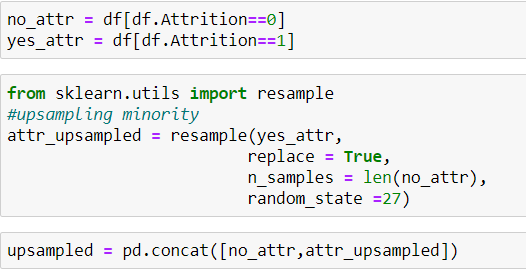
The value count for no was 1233 and for yes was 237.

In order to balance the class, we either have to down sample the majority or up sample the minority.

Here minority is yes and majority is no.

If we down sample the majority then we there will be more data loss which would not be good for the dataset.

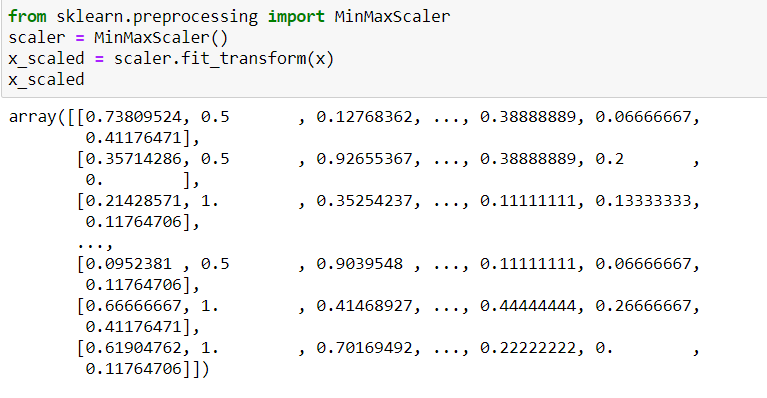
So, I have up sampled the minority ‘yes’ as shown below.



**4.4 Min Max Scaler**

Scaling of features shrinks all the value to a minimum range; this is required as the algorithms perform better when the numerical values fall under the same range.

If we do not perform scaling then the numeric in the higher range will be given more weightage from the algorithm.



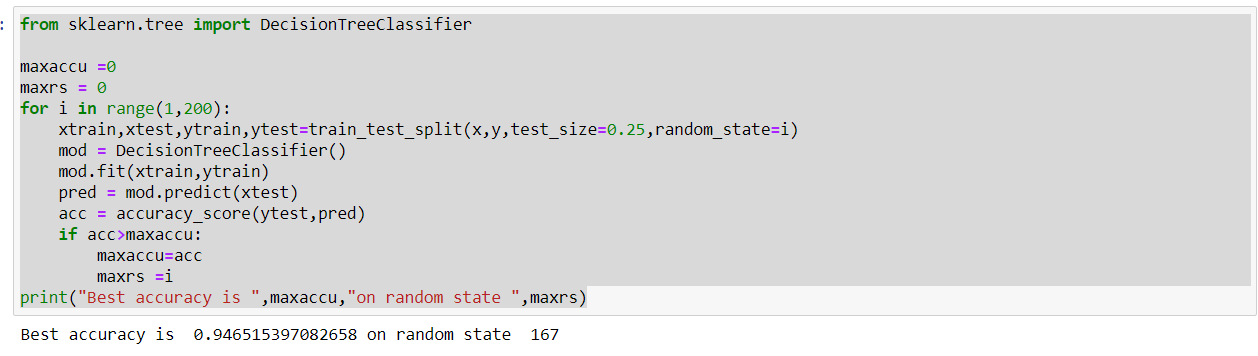
I have applied Min Max Scaler in range 0 to 1 for the features.

**4.5 Train Test Split**

We use train test split to split the given data into train data and train the model using this and the remaining as test data to check the predicted value of the model with the actual data.

I have used a for loop to iterate a model to find the best random state to get the maximum accuracy score.

The random state is to be passed in the train test split.



As seen from the above the test size is 0.25 which is that the 25% of the data will be used as the test data.

After iterating the random state is 167.



**5. Building Machine Learning Models**

Since this is a classification problem, I am using six classification algorithms

I have used the below metrics to rate the classification models and find the best model based on the metrics.

Let us understand about the metrices in brief.

Classification Accuracy is the number of appropriate predictions made as a measure of all predictions made. It is the most common test metric for classification problems.

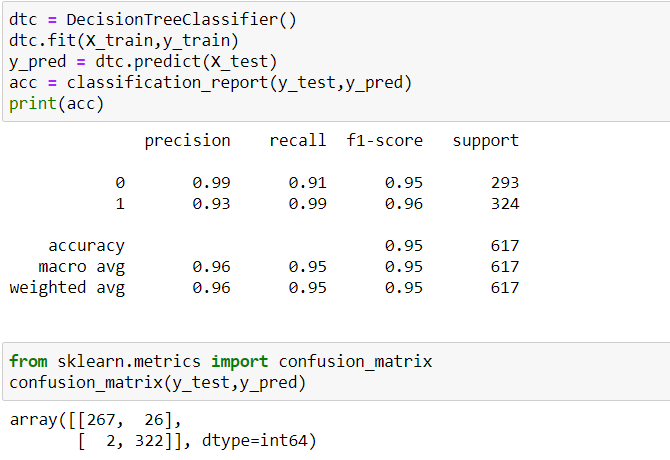
However, it is often misused as appropriate only if there is an equal number of observations in each class and all predictions and prediction errors are equally important. Not so in this project, so a different beat metric may be more appropriate.

The area under the ROC Curve (or AUC for short) is a performance metric for binary split problems. The AUC represents the model's ability to discriminate between good and bad classes, and is well suited for this task. Area 1.0 represents a model that eliminates all speculation. The 0.5 area represents a good model as random.

I will be using the classification report, auc\_roc curve and confusion matrix to get the best model.

**5.1 Decision Tree Classifier**

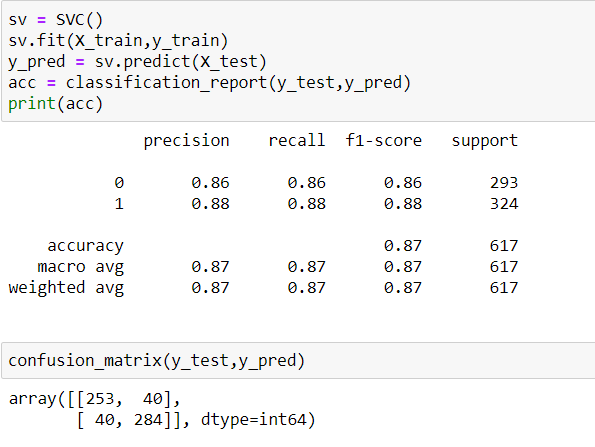
Let us take a look at Decision Tree Classifier without fine tuning the parameters.



We can see that we are getting an accuracy score of 95%

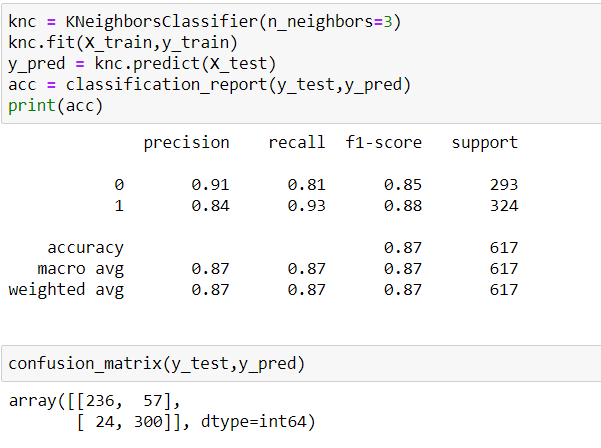
**5.2 Support Vector Classifier**

Metrics using the SVC model.



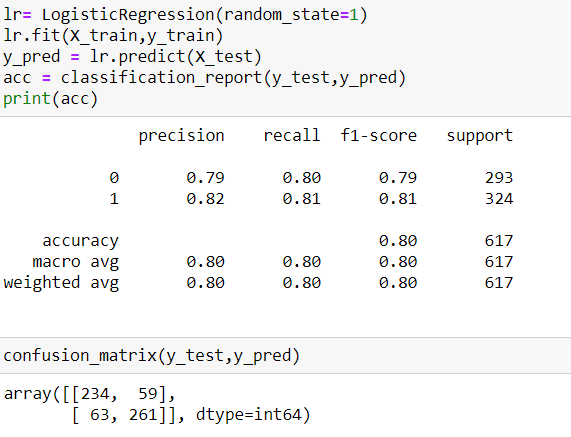
The SVC model is giving an accuracy of 87 which is less compared to DTC.

**5.3 K neighbours Classifier**



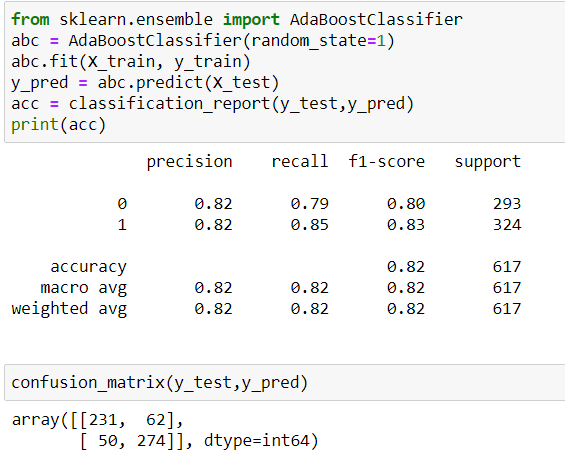
We can see that the KNC is also giving an accuracy of 87% which is less than DTC.

**5.4 Logistic Regression**



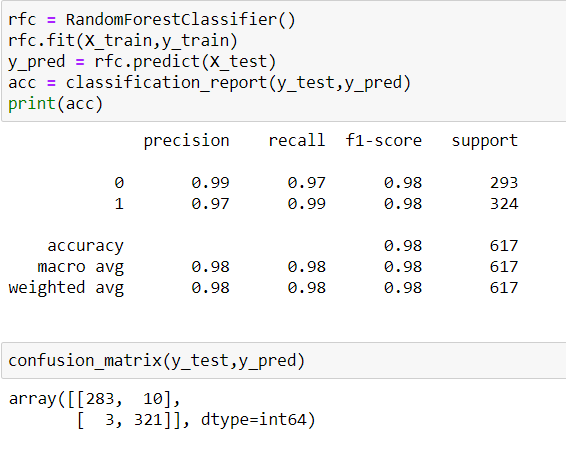
We can see that the accuracy for logistic regression is even lower at 80%.

**5.5 Ada Boost Classifier**



The ABC is giving accuracy at 82%.

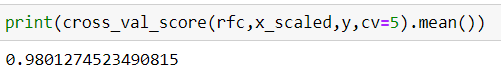
**5.6 Random Forest Classifier**



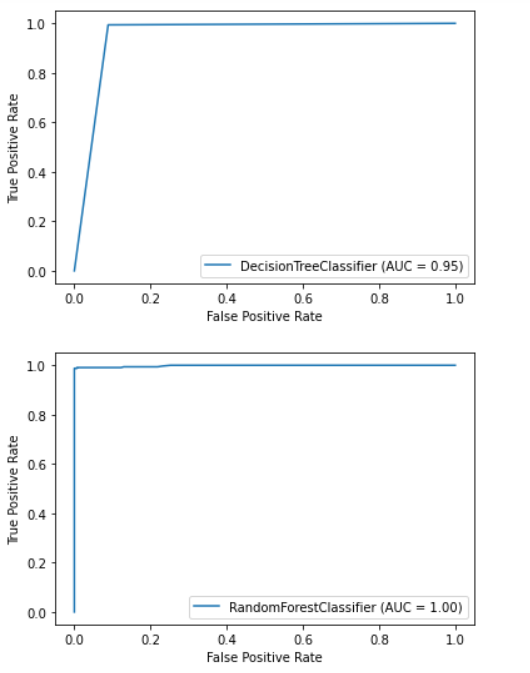
We can see that the Random Forest algorithm is giving the best accuracy at 98% and this is the best model we have now.

The precision is 0.99 with recall 0.97 and f1 score is 0.98.

I have done cv score for all the models to check with different splits of input data to the model and check if it gives the same results. I have got the best result for random forest classifier.

From all the above report, confusion matrix and cross val score we can tell that the Random Forest Regressor is the best model.

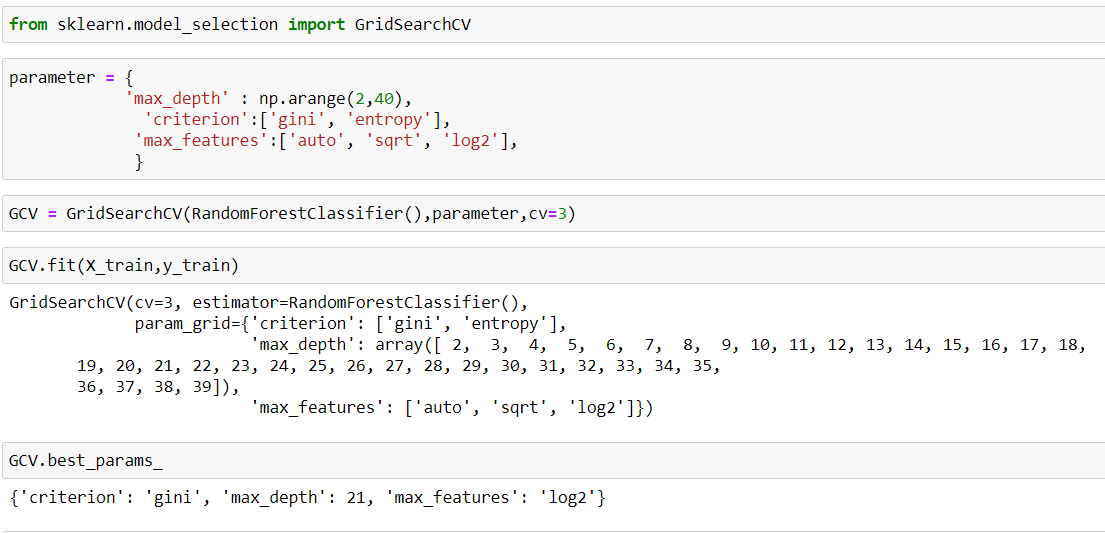
Let us however check the auc\_roc for the model.



We can see that the Random Forest Classifier as ACU score of 1.00 which is the best score.

**Grid Search CV:**

I have however, done hyperparameter tuning to improve the model.



This gives the most optimistic algorithm.

**6. Concluding Remarks**

Monthly income: the highest paid people are less likely to leave the company. Therefore, efforts should be made to collect data on industry benchmarks in the current local market to determine whether the company offers competitive wages.

Over time: overtime employees may have left the company. Efforts should therefore be made to prioritize projects with adequate funding and staffing to reduce overtime.

Age: Employees in brackets younger than 25–35-year-olds are more likely to leave. Therefore, efforts should be made to clearly articulate the long-term vision of the company and young employees equal to that vision, as well as to provide benefits in the form of explicit promotions for example.

DistanceFromHome: Employees living far from home are more likely to leave the company. Therefore, efforts should be made to provide corporate transport support for groups of employees from the same area, or through the Transformation Allowance. Initial assessment of employees based on their place of residence is probably not recommended as it can be considered a form of discrimination as long as employees make them work on time every day.

TotalWorkingYears: Employees with limited experience are less likely to leave. Employees with experience between the ages of 5 and 8 should be considered high risk for travel.

YearsAtCompany: Honest companies are less likely to leave. Employees who have spent two years celebrating years should be considered high risk of travel.

The conclusions are based only on the above dataset used.

The algorithm can be further retrained to give the optimum results, if there are more such datasets provided with more information.

Thank you for reading through and I hope you enjoyed reading it as much as I did writing it.